Semantic Classification of Nouns

Rita Policarpo

IST – Instituto Superior Técnico
L^2F – Spoken Language Systems Laboratory – INESC ID Lisboa
Rua Alves Redol 9, 1000-029 Lisboa, Portugal
rita.policarpo@tecnico.ulisboa.pt

Abstract. This paper presents several methods that can be used to achieve the semantic classification of nouns and tests the applicability of one of these algorithms through the study of the quality of the results.
Using a machine learning technique, co-training, we expect the system to increase the number of portuguese nouns semantically correctly classified in the STRING system. This algorithm receives as input data a set of previously classified nouns (seeds), which are labeled according to an existing set of semantic categories, and it performs an extensive search on a corpus looking for new sentences containing such seed-words, and also compares these sentences with the remaining sentences in the corpus in order to extract other words that fit the same word-seed context, by analyzing the syntax of the constitution of the sentences, organized as dependencies. This way, conclusions arise about new nouns that must receive as classification label the semantic category of the word seed with which they resemble in terms of the sentence context.
The predictions of the algorithm are then be submitted as proposals for the classification of a set of nouns, which are subject to approval by a human user about its correctness, allowing the expansion of the database if approved. A graphic interface was developed to present proposals for classification and allow the user to judge about their correction.
The algorithm proven to suit the classification task subject of this work. However, it is greatly influenced by the dimension of the corpus, due to its heavy comparing nature, but also by the source of the corpus, in terms of nature of the texts contained in the corpus.

1 Introduction

This work focus on the semantic classification of nouns, aiming to expand the existing database of Portuguese semantically classified nouns using one of the appropriate machine learning algorithms existing: co-training.

Machine learning algorithms can be organized in two different types: supervised and unsupervised. These learning algorithms are distinguished and applied according to the type of input available during the training phase of the algorithm:

– Supervised – These methods use labelled training examples, i.e., input where the desired output is known;

– Unsupervised – These methods use unlabeled training examples, i.e., input where the desired output is unknown.

To these, a third intermediate method can be added:

– Semi-supervised – These methods combine both labelled and unlabeled examples to generate an appropriate function or classifier.

According to these definitions, we concluded that a semi-supervised learning algorithm would be appropriate to solve the classification problem, due to the existence of a small set of labelled nouns versus a much larger set of unlabeled nouns, being this difference in the size of these sets the reason why the Co-Training algorithm (Blum and Mitchell, 1998) was chosen.

The application of the Co-Training algorithm was based on the use of training data selected from the set of 5.000 already classified Portuguese nouns and the respective contexts of those nouns obtained from the CETEMPúblico corpus (Rocha and Santos, 2000).

To understand the value of classifying nouns according to semantic tags, one must have in mind two essential concepts that are the basis of this work: syntax and semantics. While semantics is the discipline that studies the meaning of words, syntax is the discipline that studies the rules governing the formation of sentences in (natural) languages. Thus, is the part of the grammar that studies the arrangement of words in a phrase, considering their logical relation among the many possible combinations, and the different meanings that can be derived from each combination. Both these disciplines are essential to understand another concept: lexical semantics.

Lexical semantics is the study of how and what the words of a language denote (Pustejovsky, 1998). The units of meaning in lexical semantics are lexical units. Thus, words can be categorized as concepts representing different kinds of entities, using categories for those lexical units, like "Person", "Organization" or "Location", among others - these categories correspond to semantic tags. The meanings of these lexical units come from the words' individuality but also from how they relate with other linguistic elements, such as how these words relate to other words, phrases, symbols and punctuation, thus it is established by looking at its neighborhood, as in, by looking at the other words that occur in the sentence.

The studies on lexical semantics are useful to solve the problem of Word Sense Disambiguation (WSD) (Lee and Ng, 2002). Word Sense Disambiguation is the task responsible for selecting the appropriate sense (meaning) to a given word in a context, where different senses potentially attributable to that word exist. These senses can be seen as the labels of a classification problem. Thus, machine learning (ML) is a natural method to solve this problem.

One of the research areas of Natural Language Processing (NLP) is related to the human-computer interaction, which is, enabling computers to derive meaning from human or natural language input. NLP tries to develop software that works with voice recognition systems varying from search engines, speech recognition applications, automatic generators of summaries, spellcheckers, among

many others. For such applications, applying WSD is fundamental to solve lexical ambiguity, which is the existence of polysemous words that can express completely different things and whose appropriate sense in a given sentence is only distinguishable by the context analysis.

Although this work does not intend to solve WSD, the task of nouns semantic classification can be seen as a sub-problem of WSD, because it classifies nouns with tags representing all the possible attributable senses, supplying the contents for the disambiguation task. Section 2 presents a brief description on the overall architecture of the system developped to achieve semantic classification of nouns using the Co-Training algorithm.

In Section 3 the concrete results of the testing cases executed with the program implemented for this classification task are presented, and in Section 4), we discuss these results, and present the most interesting conclusions arising from this project, along with sugestions for future works that would not only improve the results of this concrete implementation, but lead to an increse of the knowledge available about this kind of classification task.

2 Architecture of the Solution

The solution implemented to this classification task takes advantage of a set of tools available at INESC-ID, thus it can be considered to be sustained by the following set of components:

- The STRING system (Mamede et al., 2012);
- The set of features of the XIP module (Ait-Mokhtar et al., 2002) (morphological, syntactic and semantic labels);
- The set of 5.000 already classified portuguese nouns and their contexts obtained from a corpus;
- A co-training algorithm implementation, for the automatic learning process of this work.

The solution implemented follows a machine learning approach to a classification problem based on (Bird et al., 2009), and presented in Figure 1.

The input of the system are the noun-context pairs, obtained from the set of already classfied nouns and the CETEMPúblico corpus (Rocha and Santos, 2000) and processed using the STRING system (Mamede et al., 2012). We use the set of features of XIP and developped a co-training implementation that presents a cyclic execution, because the algorithm uses the predictions for unlabeled data as new seeds that are added to the set and used to iteratively construct additional labelled training data.

The algorithm receives as input a set of nouns, constituted by elements of the classified nouns, named seeds. With these seeds, the first classifier of the algorithm searches for sentences containing these seeds, thus forming noun-context pairs. These noun-context pairs are fed to the second classifier of the algorithm that is responsible for gathering another contexts that match the seed ones, and then collecting the seed words that are in these new contexts, feeding the first

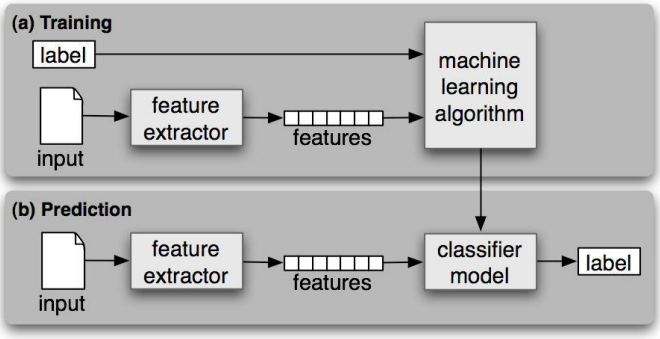

Fig. 1. Training in a classification problem (Bird et al., 2009).

classifier with the new nouns. This cycle repeats until no new information is acquired.

3 Results

With the purpose of prooving that the corpus source influences the results obtained, mainly because some semantic categories have a poor representativeness in texts of journalistic nature, thus are less probable to be found in the corpus, we opted to explore some completely opposite categories, in terms of the relationship between their semantic meaning and the content of the journalistic corpus but also their representativeness in the set of already classified nouns. Some of the categories are highly probable to be found in the corpus, while others are less probable to be found. Our choices were:

- SEM-ACT-CRIME – corresponds to criminal acts, which we considered to be highly probable to be found in journalistic corpus;
- SEM-CC-STONE – corresponds to stones or stone-sized round objects, like stone, ammonite, brick, diamond, etc.
- SPORTS - corresponds to sporting events, which we considered to be reasonably probable to be found in journalistic text;
- SEM-TOOL-MUS - corresponds to musical tools (instruments), which we considered to have low probability to be found in the corpus.

In order to determine the influence of the split factor, which reflects the division we applied to the initial corpus, in the results obtained we tested the same cases using different split factors.

All the experiments had the following configuration parameters:

- Semantic tag SEM-ACT-CRIME;
- CETEMPúblico corpus;
- Split factor: either 5 (default) or 10;
- Initial seeds not filtered;
- Final results not filtered;

3.1 SEM-ACT-CRIME (SPLIT = 5 and SPLIT = 10)

The application identified a total of 8268 words with the category SEM-ACT-CRIME, which after careful observation revealed the following partition of the results:

- 6 words identified as SEM-ACT-CRIME but already classified in our set, thus do not provide new information. They are (along with the number of occurrences):

  - Crime (8)
  - Aborto (6)
  - Dano (6)
  - Evasão (1)
  - Corrupção (2)
  - Ameaça (3)

- 3729 words identified as SEM-ACT-CRIME, but already classified in out set, and not with this tag, thus they are false positives and do not provide new information;
- 4533 words identified as SEM-ACT-CRIME, and not classified in our set, thus they are the relevant ones that can provide new information. These are the ones presented in the Results file, for human validation.

Analyzing this list, we observe that from the set of 4533 words, there are 105 new words correctly classified as SEM-ACT-CRIME, thus representing an information gain. They are (along with the number of ocurrences):

- tergiversação (3)
- logro (5)
- caciquismo (2)
- coscuvilhice (2)
- crucificação (3)
- falsidade (8)
- viciação (4)
- subordinação (2)
- adultério (1)
- cábula (2)
- escoriação (1)
- drogado (1)
- tirania (9)
- gângster (3)
- prejuízo (2)
- cabala (2)
- chicotada (3)
- incesto (3)
- compulsão (8)
- carnificina (11)
- borlista (6)
- atrito (2)
- litigação (2)
- demérito (7)
- amarração (1)
- nepotismo (1)
- laicidade (2)
- desnazificação (1)
- estocada (5)
- cobardia (8)
- poligamia (1)
- rebate (1)
- patife (2)
- leviandade (7)
- dopagem (5)
- mafioso (1)
- obscurantismo (5)
- dominação (1)
- estultícia (2)
- subterfúgio (8)
- gueto (1)
- emboscada (6)
- estratagema (3)
- prevaricador (2)
- racista (34)
- desavença (1)
- apartheid (3)
- pecador (2)
- inépcia (3)
- verdugo (4)
- empalador (1)
- perversidade (8)
- bruxaria (3)
- trafulhice (3)
- agiotagem (2)
- separatismo (1)
- gatuno (3)
- gangue (6)
- detrator (4)
- canibal (2)
- machadada (9)
- alfinetada (3)
- fuzilamento (2)
- infâmia (5)
- sanha (4)
- desonra (2)
- matracagem (1)
- aniquilação (5)
- incúria (9)

- briga (1)
- radicalista (2)
- xenófobo (2)
- conspirador (2)
- desacato (5)
- bofetada (4)
- estupefação (1)
- ultraje (1)
- tarado (2)
- processo-crime (2)
- manipulador (4)
- cruxificado (7)
- álibi (14)
- codícia (6)
- estrangulador (3)
- boémio (2)
- desertor (2)
- maledicência (5)
- enforcamento (1)
- selvajaria (5)
- vândalo (1)
- cacetada (1)
- taliban (6)
- diabolização (3)
- sevícia (1)
- mutilação (2)
- bastonada (2)
- profano (14)
- maldade (5)
- aldrabice (3)
- ditatorialização (7)
- vagabundagem (3)
- ilicitude (2)
- masoquismo (2)
- patrulhamento (1)
- devedor (6)

The 105 correct identifications from the set of 4533 makes a success percentage of 2,32%.

If we consider the total set of 8268 words, the algorithm successfully identified 111 words, thus making a success percentage of 1,34%.

The initial seeds (not filtered) found by the application for the semantic tag SEM-ACT-CRIME, totalized 81 words. Thus, it confirms the good representativeness of nouns semantically identified as criminal acts in this corpus, provenient from a journalistic source.

After carefull observation of the complete list of these 4533 new words, we verified the results are exactly the same with either the split factors. In this particular case, the partition of the corpus led to no difference in the results obtained.

The total runtime of the test with split factor 5 was 1d0h39m, while the split factor 10 had a runtime of approximately 1d3h.

3.2 SEM-CC-STONE (SPLIT = 5)

The application identified a total of 5573 words with the category sem-cc-stone, which after careful observation revealed the following partition of the results:

- 1 word (pedra), with 6 occurrences, identified as SEM-CC-STONE but already classified in our set, thus do not provide new information;
- 3108 words identified as SEM-CC-STONE, but already classified in out set, and not with this tag, thus they are false positives and do not provide new information;
- 2464 words identified as SEM-CC-STONE, and not classified in our set, thus they are the relevant ones that can provide new information. These are the ones presented in the Results file, for human validation.

Analyzing the resulting list, we observed that from the set of 2464 words, there are 8 new words correctly classified as SEM-CC-STONE, thus representing an information gain. They are (along with the number of ocurrences):

- gravilha (1)
- pedregulho (1)
- topázio (2)
- xisto (1)
- sedimento(4)
- lápis-lazúli (1)
- mineral(4)
- pedrada (7)

The 8 correct identifications from the set of 2464 makes a success percentage of 0,32%. If we consider the total set of 5573 words, the algorithm successfully identified 9 words, thus making a success percentage of 0,16%.

The initial seeds (not filtered) found by the application for the semantic tag SEM-CC-STONE, totalized 11 words. Thus, it confirms the poor representativeness of nouns semantically identified as stones/rocks in this corpus, provenient from a journalistic source.

The total runtime of this test was 7h38m.

3.3 SEM-CC-STONE (SPLIT = 10)

The application identified a total of 12116 words with the category SEM-CC-STONE, which after careful observation revealed the following partition of the results:

- 1 word identified as SEM-CC-STONE but already classified in our set, thus do not provide new information. It is (along with the number of occurrences):
  - pedra (15)
- 4233 words identified as SEM-CC-STONE, but already classified in out set, and not with this tag, thus they are false positives and do not provide new information;
- 7882 words identified as SEM-CC-STONE, and not classified in our set, thus they are the relevant ones that can provide new information. These are the ones presented in the Results file, for human validation.

Analyzing this list, we observe that from the set of 7882 words, there are 22 new words correctly classified as SEM-CC-STONE, thus representing an information gain. They are (along with the number of ocurrences):

- gesso (2)
- lancil (3)
- sedimento (11)
- minério (1)
- mineralização (2)
- rochedo (16)
- apedrejamento (8)
- sarcófago (4)
- taipa (1)
- jazida (2)
- pedras-chave (3)
- pedra-chave (2)
- meteorito (3)
- calcário (1)
- âmbar (1)
- pedrouço (1)
- pedra-de-toque (7)
- ardósia (1)
- mineral (7)
- pedrada (17)
- escombro (5)
- pedregulho (7)

The 22 correct identifications from the set of 7882 makes a success percentage of 0,28%. If we consider the total set of 12116 words, the algorithm successfully identified 23 words, thus making a success percentage of 0,19%.

The initial seeds (not filtered) found by the application for the semantic tag SEM-CC-STONE, totalized 11 words. Thus, it confirms the poor representativeness of nouns semantically identified as stones/rocks in this corpus, provenient from a journalistic source.

In this case, the split factor manifested influence in the results obtained.

The total runtime of this test was approximately 12h40m.

3.4 SPORTS (SPLIT = 5 and SPLIT = 10)

The application identified a total of 6413 words with the category SPORTS, which after careful observation revealed the following partition of the results:

- 2 words identified as SPORTS but already classified in our set, thus do not provide new information. It is (along with the number of occurrences):
  - ginástica (1)
  - futebol (2)
- 3317 words identified as SPORTS, but already classified in out set, and not with this tag, thus they are false positives and do not provide new information;
- 3094 words identified as SPORTS, and not classified in our set, thus they are the relevant ones that can provide new information. These are the ones presented in the Results file, for human validation.

Analyzing this list, we observe that from the set of 3094 words, there are 14 new words correctly classified as SPORTS, thus representing an information gain. They are (along with the number of ocurrences):

- dança-teatro (3)
- crosse (3)
- supercrosse (1)
- rali (2)
- badminton (1)
- cricket (2)
- equitação (1)
- grandes-prémio (3)
- contra-relógio (10)
- F1 (2)
- dérbi (1)
- EuroLiga (1)
- sub-20 (1)
- sub-18 (3)

The 14 correct identifications from the set of 3094 makes a success percentage of 0,45%.

If we consider the total set of 6413 words, the algorithm successfully identified 16 words, thus making a success percentage of 0,25%.

The initial seeds (not filtered) found by the application for the semantic tag SPORTS, totalized 109 words. Thus, it confirms the good representativeness of nouns semantically related to sporting events in this corpus, provenient from a journalistic source.

It is worth mention that despite the poor results, many words related to sport events were detected by the system (like “penalty”, for example), but they did not fit in the category of SPORTS events.

After carefull observation of the complete list of these 3094 new words, we verified the results are exactly the same with either the split factors. In this

particular case, the partition of the corpus led to no difference in the results obtained.

The total runtime of the test with split factor 5 was 13h19m, while the split factor 10 had a runtime of approximately 12h15m.

3.5 SEM-TOOL-MUS (SPLIT = 5)

The application identified a total of 7451 words with the category SEM-TOOL-MUS, which after careful observation revealed the following partition of the results:

- 4 words identified as SEM-TOOL-MUS but already classified in our set, thus do not provide new information. It is (along with the number of occurrences):
  - instrumento (14)
  - orgão (26)
  - prato (10)
  - disco (6)
- 3593 words identified as SEM-TOOL-MUS, but already classified in our set, and not with this tag, thus they are false positives and do not provide new information;
- 3854 words identified as SEM-TOOL-MUS, and not classified in our set, thus they are the relevant ones that can provide new information. These are the ones presented in the Results file, for human validation.

Analyzing this list, we observe that from the set of 3854 words, there are 10 new words correctly classified as SEM-TOOL-MUS, thus representing an information gain. They are (along with the number of ocurrences):

- guizo (3)
- bombo (8)
- saxophone (1)
- trombone (2)
- harmónica (5)
- maraca (2)
- carrilhão (2)
- tamborzinho (2)
- harpa (1)
- campainha (5)

The 10 correct identifications from the set of 3854 makes a success percentage of 0,26%.

If we consider the total set of 7451 words, the algorithm successfully identified 14 words, thus making a success percentage of 0,19%.

The initial seeds (not filtered) found by the application for the semantic tag SEM-TOOL-MUS, totalized 22 words. Thus, it confirms the poor representativeness of nouns semantically classified as musical instruments in this corpus, provenient from a journalistic source.

After carefull observation of the complete list of these 3854 new words, we verified the results are exactly the same with either the split factors. In this particular case, the partition of the corpus led to no difference in the results obtained.

The total runtime of the test with split factor 5 was 9h12m, while the split factor 10 had a runtime of approximately 9h15m.

4 Conclusions and Future Work

This work led to an interesting, however disappointing conclusion about natural language processing tasks: the source of the corpora determines the final results.

This is visible in the fact that the representativeness of semantic category identified examples determines the number of initial seeds detected by the application and thus, words available to be used in the comparison process for matching purposes. The more seeds available, the more number of sentences containing this word gathered, thus, higher probability of matching these sentences with new sentences containing new words.

It also reflected in the success percentages obtained: In the cases where more initial seeds were available, the comparison process had more examples to be matched against, and provided more quantity of results and with better quality, finding more new classified words to enrich our sets of Portuguese classified nouns.

It is worth mention that in this cases, where more initial seeds were available, the non-matching words found as results, despite not matching the specific semantic category we were searching, some of them were somehow related to the topic in question, what demonstrates that the algorithm was able to establish a connection between contexts and a certain topic. However, the results are limited by the semantic category to apply during human validation, since one semantic category does not cover all the words related to that topic, but only a subgroup of them.

As future work, it would be interesting to test this application using different available corpus at the L2F at INESC-ID, from different nature/source, and confirm that the kind of journalistic provenience really affects the results obtained, since journalistic text proved to be a very strict on topics addressed, influencing the semantic categories of the nouns found on the corpus.

The most difficult task in any heavy-input parsing application, is achieving a reasonable balance between resources management and quality of the results, because one cannot sacrifice accuracy of results in order to achieve shortest runtimes: it would deprecate the quality of the work.

We opted to test the execution of the algorithm with two different split factors in order to evaluate the difference in the results obtained by both, and detect how it influences the quantity and quality of the results. However, we tested not so distant values in order to evaluate the difference that a small augment or diminution on the split value causes to the results and success rates, because we believe it effectively affects them, but we want to measure how much it does with a small difference in the value, rather than testing extreme values and conclude that it affects, but now knowing exactly how much for a small portion.

On what concerns this study, we confirmed that the split factor influences the results, because despite it not revealing any differences for some of the semantic tags, it is most likely an exception, and is determined by the small difference in the values we used for testing, because the opposite occurred for the semantic tag SEM-CC-STONE, where the results were clearly different, and the test case with higher split factor presented more seed words found in the end, however

lost some of the seeds detected with the lower split factor. We used a small difference between both split factors and it revealed such an important influence, thus if we were using more distant values, the consequences and differences in the results could be dramatic. It is worth notice that this rare situation, where the split factor did not affect the results obtained despite the split factor used (not forgetting we used relatively close values) occurred for the test cases that had the more probability to be found in the corpus, and it reflected by the largest set of initial seeds. Thus, this is related to the fact that by having a good representativeness of the semantic tag in the corpus, the algorithm benefits with more examples for the comparison process, thus the probability of getting matches is also higher than in the low representativeness case, especially for the first iterations. The low representativeness cases are more affected by the corpus content because in this case the algorithm will depend on the matches of contexts, during the later iterations, starting from a small set of seeds the initial contexts are also rare. This will result in results in lower quantity and less quality.

Another topic for further investigation would be finding the better split factor to apply to the corpus, by actually quantifying the results obtained and its accuracy, studying the fluctuation of the results according to the increasing or decreasing the split factor, and documenting the degree of proportional or inversely proportional relation between these factors.

The co-training algorithm has proven to suit this kind of task, because we had a smaller set of classified nouns for different semantic tags, and a much wider set of non-classified nouns, contained in a very large corpus and, despite being influenced by the source of the corpus, it lead to a small percentage of successful results. However, it also revealed to be a really heavy algorithm to apply on corpus of such dimensions, because it implies very heavy comparisons on each iteration, forcing the corpus to be read and filtered many times, what degrades the runtime and load balance of the system.

The split factor has proven to also influence the performance of the execution of the application, because tests have revealed a difference in the runtimes with the same configuration parameters, only differing in the split factor. This is particularly visible in cases where the semantic tag is poorly represented in the corpus, like the SEM-CC-STONE tag. It would be interesting to have the same work implemented with different learning techniques, chosen from the State of Art, and compare the resources consumption of those implementations versus the long runtimes of this Co-Training algorithm implementation. Also, the quality, in terms of percentage of success, of the results obtained with different learning algorithms would be relevant to figure out what is the best algorithm to use for a semantic classification task, concerning a corpus with similar dimension to the CETEMPúblico used in this project.

Bibliography

Ait-Mokhtar, S., Chanod, J.-P., and Roux, C. (2002). Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, 8(2-3):121–144.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 92–100, Madison, Wisconsin, USA. ACM.

Lee, Y. K. and Ng, H. T. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 41–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mamede, N. J., Baptista, J., Diniz, C., and Cabarrão, V. (2012). STRING: An Hybrid Statistical and Rule-based Natural Language Processing Chain for Portuguese. *International Conference on Computational Processing of Portuguese (PROPOR 2012)*, Demo Session.

Pustejovsky, J. (1998). *The Generative Lexicon*. Bradford Books. MIT Press.

Rocha, P. A. and Santos, D. (2000). Cetempúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In das Graças Volpe Nunes, M., editor, *Actas do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR 2000)*, pages 131–140, São Paulo.